DOCUMENT RESUME

ED 419 843 TM 028 419

AUTHOR Koopmans, Matthijs

TITLE Application of HLM to the Evaluation of the Effects of Basic Skills Programs on Math Achievement.

PUB DATE 1998-04-00

NOTE 11p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Diego, CA, April 13-17, 1998).

PUB TYPE Reports - Evaluative (142) -- Speeches/Meeting Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.

DESCRIPTORS *Basic Skills; Elementary Secondary Education; *Evaluation Methods; Mathematical Models; *Mathematics Achievement; Pretests Posttests; Program Evaluation; *Regression (Statistics)

IDENTIFIERS *Hierarchical Linear Modeling; *Newark School System NJ; Stanford Achievement Tests

ABSTRACT

This paper discusses the application of Hierarchical Linear Modeling (HLM) to the evaluation of the effectiveness of Basic Skills programs in the Newark (New Jersey) school district. Whether HLM is able to handle regression effects in Basic Skills examination data is studied. The analysis uses data from the Stanford Achievement Tests for mathematics for grade 8 for 3 years (7,738 students). A comparison between HLM and a conventional pre- versus posttest design indicates that the size of the correlation between parameter estimates is considerably smaller in HLM, and not consistently in the predicted direction. It is concluded that while HLM is better able to contain regression effects, those effects remain a source of concern in the interpretation of basic skills evaluation data. (Contains one table and five references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 419 843

Application of HLM to the evaluation of the effects of Basic Skills programs on math achievement

Matthijs Koopmans
(York College/CUNY & Newark Public Schools)

Paper presented at the Annual Meeting of the American Educational Research Association, San Diego, CA, April, 1998. **Address correspondence to Matthijs Koopmans; 520 West 114 Street, Apt. 54; New York, NY 10025, or to Mkoopmans@aol.com.**

TM028419

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
✓ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Matthijs Koopmans

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

Abstract

Purpose of this paper is to discuss the application of Hierarchical Linear Modeling (HLM) to the evaluation of the effectiveness of Basic Skills programs in the district of Newark, NJ. It is examined whether HLM is better able to handle regression effects in Basic Skills evaluation data. A comparison between HLM and a conventional pre- vs. posttest design indicates that the size of the correlation between parameter estimates is considerably smaller in HLM, and not consistently in the predicted direction. It is concluded that while HLM is better able to contain regression effects, those effects remain a source of concern in the interpretation of basic skills evaluation data.

Introduction

When we evaluate the effectiveness of programs in terms of how well they enable the students in the district to learn and grow, it is implied that we are interested in the assessment of change. By their very nature, Title I programs intend to enhance student learning, and improve opportunities for those students who have been educationally deprived. The two objectives which guide program evaluations, such as Title I in school districts such as the Newark Public Schools are (1) determination whether students in the benchmark grades (4^{th}, 8^{th}, and 11^{th} grades) meet certain state mandated content and performance standards, and (2) a determination of whether students in the other grades make adequate progress towards attainment of those standards. Evaluations which are concerned with these two objectives address questions which are inherently longitudinal: progress takes place over time, and meeting high standards results from a learning process.

Over the years, the justification of Title I and other remedial programs has usually been reflective of a longitudinal conceptualization of student achievement, while the methodological approach to the evaluation of student progress has typically been cross sectional. The purpose of this paper is to discuss how recently developed longitudinal approaches can be applied to the evaluation of program and school effectiveness in the district, and how they can remedy some of the shortcomings of traditional techniques.

Traditional Evaluation Models

The traditional Title I evaluation consists of two components, referred to as the Model A Evaluation, and the Model C Evaluation. The Model A evaluation refers to the comparison of pretest scores to posttest scores in order to determine whether progress has been made in a given year relative to the previous year. Such a comparison can be carried out either with, or without, an adjustment of the pretest averages for regression effects. Regression effects have complicated the interpretation of the findings using a Model A design, since those who scored lower in a given year tend to have higher scores in a subsequent year, whereas those who score at the higher end of the continuum tend to score lower in the subsequent year. Since the determination of eligibility for Title I services is based on those very same test scores, results of a comparison between Title I students and Non-Title I students predictably shows improvement in the Title I group, and a decline in the non-Title I group.

The Model C evaluation consists of three steps. (1) on the basis of test scores of the previous year, it is determined whether students are eligible for Title I services. Students scoring below a given cut-off point are determined eligible for Title I services, and those scoring at or above that cut off are not considered eligible. (2) A line is fitted, which regresses posttest scores on the pretest scores of students who were *not* in Title I at a given year. An expected average posttest score is then computed for those students who were in Title I. The accuracy of this estimate depends on the size of the correlation between the pre, and posttest scores. With low correlations, the expected average for the Title I group tends to be overestimated. Since correlations between pretest and posttest tend to be modest in the population of this district, an adjustment of the estimation was deemed necessary: (3) To address this concern, a 99%

confidence interval is constructed around the estimated posttest mean for Title I students. The lower boundary of the interval is used as an estimated mean for the Title I students. The posttest score averages by grade and school are then compared to their estimated mean to determine whether the program exceeded expectations (Madhere, n.d.).

Two methodological problems in traditional evaluation designs

The evaluation of Title I programs in Newark and elsewhere, has typically been plagued by two methodological problems. First, our ability to detect effects of the program is frustrated by regression effects, which are inherent to pre- vs. posttest comparisons. Those students, who score extremely low in the pretest year, will be in the basic skills group, whereas students who score extremely high in the pretest year will be in the comparison group. When the two groups are compared one year later, students with extreme scores in the previous year will tend to have scores, which are less extreme, and averages in the eligible as well as in the ineligible group will both tend toward the mean. Consequently, there will be an increase in the scores of students receiving Title I services, regardless of the merits of those services. Capitalization of this effect in our evaluation practices creates an unduly favorable picture of the effectiveness of Title I programs. Evaluators have long been aware of this issue, and various methods have been proposed to address the problem (e.g., repeated measures analysis of variance, partial correlations, and various other types of adjustment). The purpose of this paper, is to illustrate that growth modeling within an HLM framework (see Bryk & Raudenbush, 1992; Willett, 1990) better contains these regression effects than models based on traditional pre- vs. posttest designs.

The second methodological problem in traditional evaluation models has been our inability to deal with student- and school level characteristics at the same time. It is important to distinguish these levels of information when determining the effectiveness of their Title I programs. One needs to be able to adjust student level estimates, such as program enrollment status, for sources of variation on the school level, variables such as student – teacher ratio, percentage of students in bilingual programs, and mobility rate of the student body. These variables are often beyond the control of the schools (Webster, et al., 1996). The analysis presented in this paper includes school level variables as covariates in the comparisons of student performance over time in Basic Skills programs vs. non-Basic Skills programs.

Procedure

The present analysis compares the results of a traditional pre- vs. postest design to those of Hierarchical Linear Modeling using the same achievement data. This analysis is specifically concerned with math achievement, as measured by the Stanford-8 Achievement Test, in 1994, 1995, and 1996. Only students were included in the analysis for whom data could be obtained for all three years. Thus, 7738 students were included. The analysis was concerned with grades 3, 5, 6, and 8. Demographic characteristics of the student included in this analysis reflect those of the district of Newark at large. Of the 90% minority students in the district, about two thirds, is Black or African American, and the majority of the remaining one third are Hispanic (with Spanish or Portuguese as the native language

The statistical analysis of data consisted of two parts. A traditional pre- posttest design was used to compare the average math NCE scores in 1995 to those in 1996, and gainscores (i.e., the difference between the 1995 and 1996 NCE scores) were computed for each grade level. This part of the analysis corresponds to the traditional Model A design. Correlations between size of the 1995-1996 gain and the 1995 scores were computed to estimate the magnitude of the regression effect.

The following HLM models were tested (terminology is used in the conventional manner, see Raudenbush & Bryk, 1990; Bryk, Raudenbush & Congdon, 1996):

Level 1 (the within subjects growth model):

$$\hat{Y} = \pi_0 + \pi_1 \text{ year} + e$$

Where $\hat{Y}$ stand for the NCE math scores. π_0, the intercept, stands for the predicted entry level, i.e., the predicted 1995 scores for each student, based on their own 1994 – 1996 growth trajectory. π_1 stands for the within subjects growth rate over the three-year period, and e is an error term.

Level 2 (the student level between subjects model):

$$\pi_0 = \beta_{00} + \beta_{01}\text{Basic Skills Group} + R_0$$

$$\pi_1 = \beta_{10} + \beta_{11}\text{Basic Skills Group} + R_1$$

The predictor in this model is membership in the basic skills group between 1995 and 1996. β_{00} and β_{10} are intercepts. β_{01} and β_{11} are the parameter estimates for the relative effects of basic skills group membership on, respectively, the intercept of the growth model, and on the individual growth rate. R_0 and R_1 are error terms.

Level 3 (the between schools model):

$$\beta_{00} = \gamma_{000} + \gamma_{001}\text{BILPCT} + \gamma_{002}\text{CH1PCT} + U_0$$

$$\beta_{01} = \gamma_{010} + U_1$$

$$\beta_{10} = \gamma_{100} + \gamma_{101}\text{BILPCT} + \gamma_{102}\text{CH1PCT} + U_2$$

$$\beta_{11} = \gamma_{110} + U_3$$

In this model, BILPCT stands for the percentage of students enrolled in the bilingual program within each school, and CH1PCT stands for the percentage of students enrolled in basic skills programs within each school. A stepwise model selection at the school level (see Bryk and

Raudenbush, 1992) indicated that inclusion of these variables significantly affected the estimation of the level 2 intercepts. In the present analysis, the level 3 predictors should be interpreted as covariates.

In the district of Newark, the percentage of students enrolled in basic skill programs is confounded with the school level SES indicator. CH1PCT was used because it turned out to be a somewhat more reliable indicator.

Results

Table 1 compares the results of the HLM analysis to those of a traditional pre- vs. posttest design. The HLM estimates reported are the average predicted entry level (π_0) for each grade, the slope of the average growth trajectory for each grade (π_1), and the correlation between the predicted mean entry level (π_0), and the predicted mean growth rate (π_1) for each grade. Each of these estimates is adjusted for the level 3 predictors mentioned above. The traditional estimates reported are the observed 1995 NCE math scores for each grade level, the 1995-1996 gain scores, and the correlation between the 1995 scores and the gainscores.

It can be seen in the table that predicted entry levels are higher than the observed 1995 scores in the basic skills group, and lower in the non-basic skills group. This difference reflects the fact that the 1995 scores largely determine group membership, and that 1994 and 1996 scores therefore tend to be higher than the 1995 scores for the basic skills group, and lower for the nonbasic skills group. Both HLM and conventional estimates reveal statistically significant differences between the basic skills group and the comparison group: observed and predicted entry levels are significantly higher in the comparison group than in the basic skills group.

The table also shows the fitted slopes of the individual growth functions for each grade level. These slopes depict the yearly decline, in NCE points, in math achievement, as estimated by the growth model described above. It can be seen that there is a decline in the predicted math achievement scores in all instances, except for the comparison group in grade 8, for whom the predicted achievement goes up for .2 NCE points. A steeper decline for the basic skills group than the comparison group is observed in the fifth grade (3.7 NCE points and 2.9 points respectively). In the third grade, the decline is steeper in the comparison group than in the basic skills group (2.5 NCE points and 1.5 NCE points respectively). No statistically significant differences are observed between the two groups in grade 6.

Differences in gain scores between the basic skills and comparion groups are quite pronounced. In the third grade, they go up by 9.2 NCE points for the basic skills group, and they go down by 7.5 NCE points for the comparison group. In grade 5, similarly, there is an increase of 1.1 point for the basic skills group and a decline of 5.4 points for the comparison group. In grade 6, no difference between 1995 and 1996 is observed in the comparison group, while in the basic skills group, scores still go up by 4.4 NCE points. In grade 8, finally, differences between the two groups were not statistically significant.

Correlation measures indicate a stronger association between observed entry level scores and gainscores (the conventional estimates), than between predicted entry level and predicted growth rate (the HLM estimates). Correlations between the conventional estimates fall between -.17 (grade 8) and -.42 (grade 3), whereas the correlations between the HLM estimates fall

between +.20 and -.08. All correlations are statistically significant at the $\alpha = .05$ level (two-tailed), except for the correlation between the HLM estimates in grade 3.

Discussion

The conventional estimates used in pre- vs. posttest designs show a highly predictable pattern of results. Lower observed entry level scores (1995 NCE scores) are associated with higher gain scores, and students enrolled in basic skills programs in the subsequent year (1995 – 1996) make positive gains whereas those who are not enrolled in such programs make negative gains (i.e., their NCE scores decline). It can also be seen in Table 1 that this pattern becomes less pronounced as grade level gets higher. However, correlations remain negative and statistically significant for all grade levels. In the evaluation of program effectiveness, it is therefore difficult to determine to what extent progress in the basic skills group is 'real' and to what extent such progress is due to regression effects.

Correlations between predicted entry level and slope of the growth curves reveal a much less predictable pattern in the HLM estimates. In grade 5 and 6, these correlations are statistically significant, and negative as well, indicating that higher predicted entry levels tend to be accompanied by lower growth rates. However, the size of this effect is considerably smaller in the HLM model than in the conventional model. In the third grade, where the effect is most strongly discernible in the conventional model, it is absent in the HLM model. In the seventh and eighth grade, on the other hand, correlations between estimates are statistically significant in both models, but in opposite directions. In conventional models, a lower observed entry level tends to be followed by a higher gain score in these grades, whereas a lower predicted entry level in the HLM model tends to be associated with lower growth rates as well.

What can we attribute these differences to? Several factors are important. First, HLM models consider learning history. In the analysis presented here, growth curves are fitted to three years of data (1994 – 1996), whereas conventional estimates consider only 1995 and 1996. As a result, the overwhelming influence of the entry level scores on the basis of which basic skills group membership is determined is dampened. This makes the results of the estimation of student progress less sensitive to the effects of the extremes through which the entry level estimates are dissociated in the two groups. Second, the HLM estimates are adjusted for school level influences. In this analysis, the percentage of students in bilingual programs, and in chapter 1 programs are used as covariates.

We cannot eliminate regression effects from these data, and this analysis indicates that such effects remain a source of concern when HLM approaches are used. This analysis suggests, however, that these effects are considerably less pronounced in the two groups. Moreover, a consideration of learning history and school level information reduces sensitivity of the parameter estimates to these effects.

Aside from its ability to better contain regression effects, there is also a substantive appeal to the use of HLM. The assessment of student progress toward meeting academic standards requires a longitudinal as well as a hierarchical conception of evaluation data: A longitudinal conception because progress takes place over time, and a hierarchical conception because student performance is affected by student level characteristics, as well as school level

characteristics.

There are also drawbacks to the use of HLM for evaluation purposes, however. To control for learning history, it is necessary to use test scores for at least three time points, rather than two, as is the case in traditional pre- vs. posttest designs. With each additional year, there is attrition of cases. The advantage of considering multiple time points needs to be weighted against the loss of cases that result from it. The drawback is not specific to HLM but inherent to the use of longitudinal designs. The possibility needs to be entertained, in any event, that the attrition of cases is systematic. HLM also requires a sufficiently large number of schools that can be included in the analysis to ensure the stability of estimates.

References

Bryk, A. S. & Raudenbush, S. W. (1992) Hierarchical Linear Models: Applications and data analysis methods. Newbury Park, CA: Sage.

Bryk, A. S. Raudenbush, S. W. & Congdon, R. T. (1996) HLM: Hierarchical Linear and Nonlinear Modeling with the HLM/2L and HLM/3L programs. Chicago, IL: Scientific Software International.

Madhere, S. (n.d.) Efficiency analysis: Enhancing the statistical and evaluative power of the regression - Discontinuity design. In Engineering Educational Excellence: The View of Urban Researchers. Publication of the Board of Education, Newark, NJ.

Raudenbush, S. W. & Bryk, A. S. (1990) Methodological advances in analyzing the effects of schools and classrooms on student learning. *Review of Research in Education*, **15**, 423-475.

Willett, J. B. (1990) Questions and answers in the measurement of change. *Review of Research in Education*, **15**, 345-422.

Table 1. Results of the comparison of HLM parameter estimates with those from pretest-postest designs: Means, standard deviations, and correlations.

Grade (n students/ n schools)	HLM Estimates			Conventional Estimates (Pre- vs. Posttest design)		
Basic Skills Group	Entry (π_0)	Slope (π_1)	$r(\pi_0,\pi_1)$	Obs. Entry NCE '95	Gainscore '95-'96	r[b]
3 (1988/52)						
BS	31.7	-1.5	-.04[c]	24.5	9.2	-.42
	(9.8)	(3.3)		(11.6)	(16.3)	
NBS	56.6	-2.5		60.1	-7.5	
	(13.4)	(5.2)		(19.1)	(18.0)	
5 (2035/52)						
BS	30.3	-3.7	-.08	27.1	1.1	-.35
	(8.7)	(3.8)		(10.9)	(15.7)	
NBS	54.9	-2.9		56.5	-5.4	
	(14.1)	(3.1)		(18.3)	(15.9)	
6 (1903/41)						
BS	32.5	-1.3	-.07	28.7	4.4	-.30
	(9.1)	(1.8)		(11.4)	(14.1)	
NBS	55.6	-1.5[a]		54.6	-.0	
	(15.0)	(3.3)		(18.1)	(13.8)	
8 (1812/39)						
BS	31.3	-.3	.17	29.8	1.9	-.17
	(7.8)	(3.5)		(8.7)	(12.1)	
NBS	54.8	.2		54.9	.8[a]	
	(14.5)	(3.3)		(16.7)	(12.2)	

Note: BS: Basic Skills Group
NBS: Non Basic Skills Group

[a] Differences between Basic Skills group and Non Basic Skills group are not statistically significant at the $p < .05$ level.

[b] Correlation between 1995 NCE and 1995 – 1996 gainscores.

[c] Correlation between HLM estimates are not statistically significant at the $p < .05$ level (two-tailed).